

Joint Inference of Groups, Events and Human Roles in Aerial Videos

Tianmin Shu¹, Dan Xie¹, Brandon Rothrock², Sinisa Todorovic³ and Song-Chun Zhu¹

¹ Center for Vision, Cognition, Learning and Art, University of California, Los Angeles

{stm512, xiedan}@g.ucla.edu sczhu@stat.ucla.edu

²Jet Propulsion Laboratory, California Institute of Technology

brandon.rothrock@jpl.nasa.gov

³School of Electrical Engineering and Computer Science, Oregon State University

sinisa@onid.orst.edu

Abstract

With the advent of drones, aerial video analysis becomes increasingly important; yet, it has received scant attention in the literature. This paper addresses a new problem of parsing low-resolution aerial videos of large spatial areas, in terms of 1) grouping, 2) recognizing events and 3) assigning roles to people engaged in events. We propose a novel framework aimed at conducting joint inference of the above tasks, as reasoning about each in isolation typically fails in our setting. Given noisy tracklets of people and detections of large objects and scene surfaces (e.g., building, grass), we use a spatiotemporal AND-OR graph to drive our joint inference, using Markov Chain Monte Carlo and dynamic programming. We also introduce a new formalism of spatiotemporal templates characterizing latent sub-events. For evaluation, we have collected and released a new aerial videos dataset using a hex-rotor flying over picnic areas rich with group events. Our results demonstrate that we successfully address above inference tasks under challenging conditions.

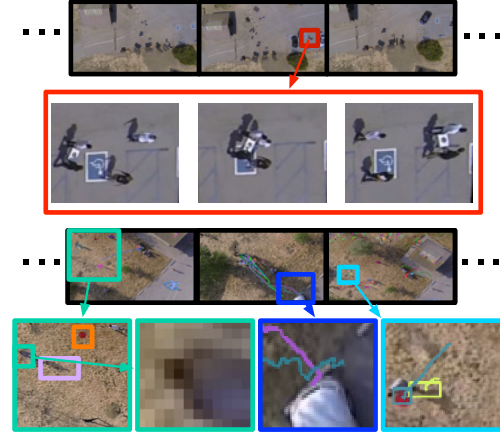


Figure 1: Our low-resolution aerial videos show top-down views of people engaged in a number of concurrent events, under camera motion. Different types of challenges are color-coded. The red box marks a zoomed-in video part with varying dynamics among people and their roles *Deliverer* and *Receiver* in *Exchange Box*. The green marks extremely low resolution and shadows. The blue indicates only partially visible *Car*. The cyan marks noisy tracking of person and the small object *Frisbee*.

1. Introduction

1.1. Motivation and Objective

Video surveillance of large spatial areas using unmanned aerial vehicles (UAVs) becomes increasingly important in a wide range of civil, military and homeland security applications. For example, identifying suspicious human activities in aerial videos has the potential of saving human lives and preventing catastrophic events. Yet, there is scant prior work on aerial video analysis [13, 12, 29], which for the most part is focused on tracking people and vehicles (with few exceptions [23]) in relatively sanitized settings.

Towards advancing aerial video understanding, this paper presents a new problem of parsing extremely low-resolution aerial videos of large spatial areas, such as picnic areas rich with co-occurring group events, viewed top-down under camera motion, as illustrated in Fig. 1 and 2. Given an aerial video, our objectives include:

1. Grouping people based on their events;
2. Recognizing events present in each group;
3. Recognizing roles of people involved in these events.

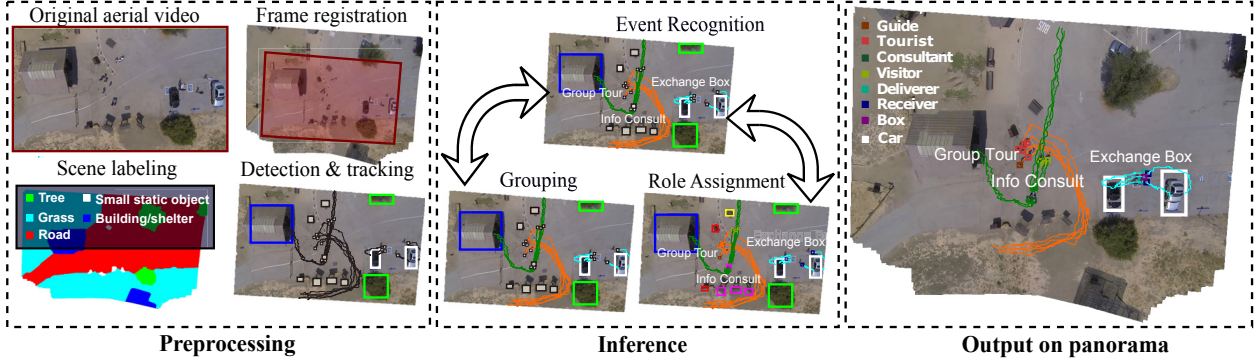


Figure 2: The main steps of our approach. Our recognition accounts for the temporal layout of latent sub-events, people’s roles within events (e.g., *Guide*, *Visitor*), and small objects that people interact with (e.g., *Box*, *trash bin*). We iteratively optimize groupings of the foreground trajectories, infer their events and human roles (color-coded tracks) within events.

1.2. Scope and Challenges

As illustrated in Fig. 1, we focus on videos of relatively wide spatial areas (e.g., parks with parking lots) with interesting terrains, taken on-board of a UAV flying at a large altitude (25m) from the ground. People in such videos are formed into groups engaged in different events, involving complex n -ary interactions among themselves (e.g., a *Guide* leading *Tourists* in *Group Tour*), as well as interactions with objects (e.g., *Play Frisbee*). Also, people play particular roles in each event (e.g., *Deliverer* and *Receiver* roles in *Exchange Box*).

1. **Low resolution.** People and their portable objects are viewed at an extremely low resolution. Typically, the size of a person is only 15×15 pixels in a frame, and small objects critical for distinguishing one event from another may not be even distinguishable by a human eye.

2. **Camera motion** makes important cues for event recognition (e.g., object like *Car*) only partially visible or even out of view, and thus may require seeing longer video footage for their reliable detection.

3. **Shadows in top view** make background subtraction very challenging.

Unfortunately, popular appearance-based approaches to detecting people and objects used to produce input for recognizing group events and interactions [25, 7, 32, 16, 30, 9] do not handle the above three challenges. Thus we have to depart from the appearance-based event recognition.

In addition, in the face of these challenges, the state of the art methods in people and vehicle tracking frequently miss to track moving foreground, and typically produce short, broken tracklets with a high rate of switched track IDs.

4. **Space-time dynamics.** Our events are characterized by both very large and very small space-time dynamics within a group of people. For example, in the event of a line forming in front of a vending machine, called *Queue*

for *Vending machine*, the participants may be initially scattered across a large spatial area, and may form the line very slowly, while partially occluding one another when closely standing in the line.

1.3. Overview of Our Approach

As Fig. 2 illustrates, our approach consists of two main steps:

1. **Preprocessing.** We ground our approach onto noisy detections and tracking. Foreground tracking under camera motion is made feasible by registering video frames onto a reference plane. By frame registration, we generate a panorama for scene labeling. Due to the challenges mentioned in Sec. 1.2, tracking of small portable objects and people produces highly unreliable frequently broken tracklets, with a high miss rate. We improve the initial tracking results by agglomeratively clustering tracklets into longer trajectories based on their spatial layout and velocity. We detect large objects (e.g. buildings, cars) using the approach of [31], and classify superpixels [1] of the panorama for scene labeling.

2. **Inference.** We seek event occurrences in the space-time patterns of the foreground trajectories and their relations with the detections of objects in the scene. To constrain our recognition hypotheses under uncertainty, we resort to domain knowledge represented by a probabilistic grammar – namely, a spatiotemporal AND-OR graph (ST-AOG). ST-AOG encodes decompositions of events into temporal sequences of sub-events. Sub-events are defined by our new formalism called *latent spatiotemporal templates* of n -ary relations among people and objects. The templates jointly encode varying spatiotemporal relations of characteristic roles of all people, as well as their interactions with objects, while engaged in the event.

We specify an iterative algorithm based on Markov Chain Monte Carlo (MCMC [15]) along with dynamic pro-

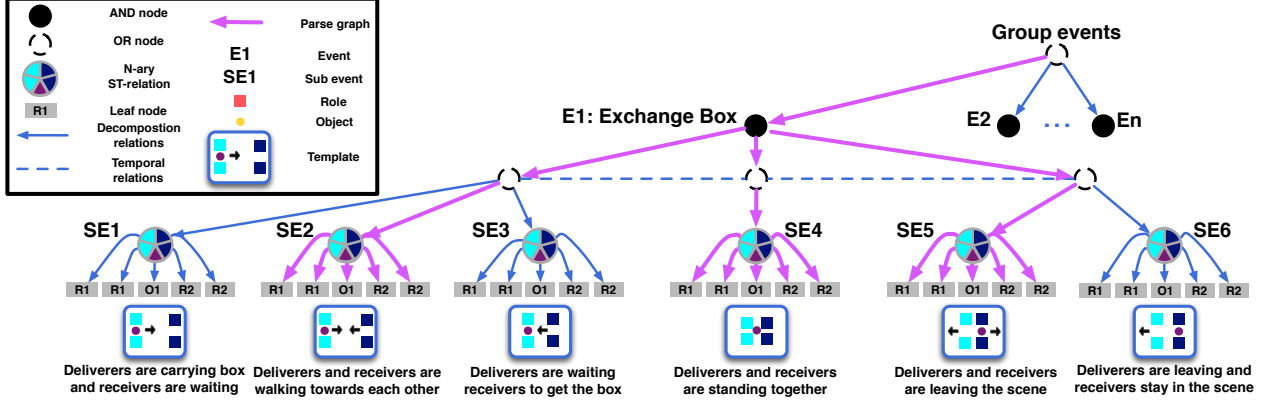


Figure 3: A part of ST-AOG for *Exchange Box*. The nodes are hierarchically connected (solid blue) into three levels, where the root level corresponds to events, middle level encodes sub-events, and leaf level is grounded onto foreground tracklets and small static objects in the video. The lateral connections (dashed blue) indicate temporal relations of sub-events. The colored pie-chart nodes represent templates of n -ary spatiotemporal relations among human roles and objects (see Fig. 4). The magenta edges indicate an inferred parse graph which recognizes and localizes temporal extents of events, sub-events, human roles and objects in the video.

gramming (DP) to jointly infer groups, events and human roles.

1.4. Prior Work and Our Contributions

Our work is related to three research streams.

Event Recognition in Aerial Videos. Prior work on aerial image and video understanding typically puts restrictions on their settings for limited tasks. For example, [27] requires robust motion segmentation and learning of object shapes for tracking objects; [12] recognizes people based on background subtraction and motion; and [29] depends on appearance-based regressor and background subtraction for tracking vehicles. Regarding the objectives, these approaches mainly focus on detecting and tracking people or vehicles [38, 23, 13]. We advance prior work by relaxing their assumptions about the setting, and by extending their objectives to jointly infer groups, events, human roles.

Group Activity Recognition. Simultaneous tracking of multiple people, discovering groups of people, and recognizing their collective activities have been addressed only in every-day videos, rather than aerial videos [8, 32, 17, 10, 18, 7, 6, 5, 34, 36]. Also, work on recognizing group activities in large spatial scenes requires high-resolution videos for a “digital zoom-in” [4]. As input, these approaches use person detections along with cues about human appearance, pose, and orientation — i.e., information that cannot be reliably extracted from our aerial videos. There are also some trajectory-based methods for event recognition [21, 35, 20], but they focus on simpler events compared to what we discuss in this paper. Regarding the representation of collective activities, prior work has used a descriptor of human locations and orientations, similar to shape-context [7, 5]. We

advance prior work with our new formalism of latent spatiotemporal template of human roles and their interactions with other actors and objects.

Recognition of Human Roles. Existing work on recognizing social roles and social interactions of people typically requires perfect tracking results [30], reliable estimation of face direction and attention in 3D space [9], detection of agent’s feet location in the scene [41], and thus are not applicable to our domain. Our approach is related to recent approaches aimed at jointly recognizing events and social roles by identifying interactions of sub-groups [10, 18, 16, 14].

Contributions:

1. Addressing a more challenging setting of aerial videos;
2. New formalism of latent spatiotemporal templates of n -ary relations among human roles and objects;
3. Efficient inference using dynamic programming aimed at grouping, recognition and localizing temporal extents of events and human roles
4. New dataset of aerial videos with per-frame annotations of people’s trajectories, object labels, roles, events and groups.

2. Representation

2.1. Representing of Group Events by ST-AOG

Similar with hierarchical representation in [11, 19, 24, 26], domain knowledge is formalized as ST-AOG, depicted in Fig. 3. Its nodes represent the following four sets of concepts: events $\Delta_E = \{E_i\}$; sub-events $\Delta_L = \{L_a\}$; human roles $\Delta_R = \{R_j\}$; small objects that people interact with $\Delta_O = \{O_j\}$; and large objects and scene surfaces

$\Delta_S = \{S_j\}$. A particular pattern of foreground trajectories observed in a given time interval gives rise to a sub-event, and a particular sequence of sub-events defines an event.

Edges of the ST-AOG represent decomposition and temporal relations in the domain. In particular, the nodes are hierarchically connected by decomposition edges into three levels, where the root level corresponds to events, middle level encodes sub-events, and leaf level is grounded onto foreground tracklets and object detections in the video. The nodes of sub-events are also laterally connected for capturing “followed-by” temporal relations of sub-events within the corresponding events.

ST-AOG has special types of nodes. An AND node, \wedge , encodes a temporal sequence of latent sub-events required to occur in the video so as to enable the event occurrence (e.g., in order to *Exchange Box*, the *Deliverers* first need to approach the *Receivers*, give the *Box* to the *Receivers*, and then leave). For a given event, an OR node, \vee , serves to encode alternative space-time patterns of distinct sub-events.

2.2. Sub-events as Latent Spatiotemporal Templates

A temporal segment of foreground trajectories corresponds to a sub-event. ST-AOG represents a sub-event as the *latent* spatiotemporal template of n -ary spatiotemporal relations among foreground trajectories within a time interval, as illustrated in Fig. 4. In particular, as an event is unfolding in the video, foreground trajectories form characteristic space-time patterns, which may not be semantically meaningful. As they frequently occur in the data, they can be robustly extracted from training videos through unsupervised clustering. Our spatiotemporal templates formalize these patterns within the Bayesian framework using unary, pairwise, and n -ary relations among the foreground trajectories. In addition, our unsupervised learning of spatiotemporal templates address unstructured events in a unified manner. Namely, more structured events need more templates and an unstructured one is represented by a single template.

Unary attributes. A foreground trajectory, $\Gamma = [\Gamma^1, \dots, \Gamma^k, \dots]$, can be viewed as spanning a number of time intervals, $\tau_k = [t_{k-1}, t_k]$, where $\Gamma^k = \Gamma(\tau_k)$. Each trajectory segment, Γ^k , is associated with unary attributes, $\phi = [r^k, s^k, c^k]$. Elements of the role indicator vector $r^k(l) = 1$ if Γ^k belongs to a person with role $l \in \Delta_R$ or object class $l \in \Delta_O$; otherwise $r^k(l) = 0$. The speed indicator $s^k = 1$ when the normalized speed of Γ^k is greater than a threshold (we use 2 pixels/sec); otherwise, $s^k = 0$. Elements of the closeness indicator vector $c^k(l) = 1$ when Γ^k is close to any of the large objects or types of surfaces detected in the scene indexed by $l \in \Delta_S$, such as *Building*, *Car*, for a threshold (70 pixels); o.w., $c^k(l) = 0$.

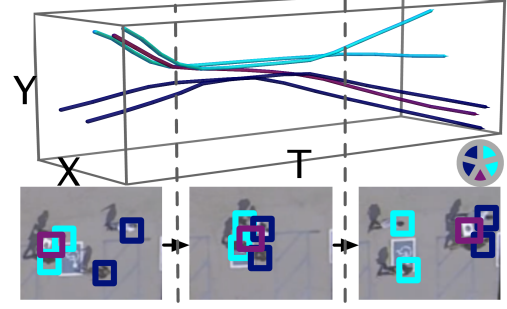


Figure 4: Three example templates of n -ary spatiotemporal relations among foreground trajectories extracted from the video (XYT-space) for the event *Exchange Box*. The recognized roles *Deliverers*, *Receivers* and the object *Box* in each template are marked cyan, blue and purple, respectively. Spatiotemporal templates are depicted as colored pie-chart nodes in Fig. 3.

Pairwise relations. of a pair of trajectory segments, Γ_j^k and $\Gamma_{j'}^k$, are aimed at capturing spatiotemporal relations of human roles or objects represented by the two trajectories, as illustrated in Fig. 4. The pairwise relations are specified as: $\phi_{jj'} = [d_{jj'}^k, \theta_{jj'}^k, r_{jj'}^k, s_{jj'}^k, c_{jj'}^k]$, where $d_{jj'}^k$ is the mean distance between Γ_j^k and $\Gamma_{j'}^k$; $\theta_{jj'}^k$ is the angle subtended between Γ_j^k and $\Gamma_{j'}^k$; and the remaining three pairwise relations check for compatibility between the aforementioned binary relations as: $r_{jj'}^k = r_j^k \oplus r_{j'}^k$, $s_{jj'}^k = s_j^k \oplus s_{j'}^k$, $c_{jj'}^k = c_j^k \oplus c_{j'}^k$, where \oplus denotes the Kronecker product.

n -ary relations. Towards encoding unique spatiotemporal patterns of a set of trajectories, we specify the following n -ary attribute. A set of trajectory segments, $G_i(\tau_k) = G_i^k = \{\Gamma_j^k\}$, can be described by a 18-bin histogram h^k of their velocity vectors. h^k counts orientations of velocities at every point along the trajectories in a polar coordinate system: 6 bins span the orientations in $[0, 2\pi]$, and 3 bins encode the locations of trajectory points relative to a given center. As the polar-coordinate origin, we use the center location of a given event in the scene.

Unsupervised Extraction of Templates. Given training videos with ground-truth partition of all their ground-truth foreground trajectories G into disjoint subsets $G = \{G_i\}$. Every G_i can be further partitioned into equal-length time intervals $G_i = \{G_i^k\} (|\tau^k| = 2\text{sec})$. We use K-means clustering to group all $\{\Gamma_{i,j}^k\}$, and then estimate spatiotemporal templates $\{L_a\}$ as representatives of the resulting clusters a . For K-means clustering, we use ground-truth values of the aforementioned unary and pairwise relations of $\{\Gamma_{i,j}^k\}$. In our setting of 11 categories of events occurring in aerial videos, we estimate $|\Delta_L| = 27$ templates.

3. Formulation and Learning of Templates

Given the spatiotemporal templates, $\Delta_L = \{L_a\}$, extracted by K-means clustering from training videos (see Sec. 2.2), we will conduct inference by seeking these latent templates in foreground trajectories of the new video. To this end, we define the log-likelihood of a set of foreground trajectories $G = \{\Gamma_j\}$ given $L_a \in \Delta_L$ as

$$\begin{aligned} \log p(G|L_a) &\propto \sum_j \mathbf{w}_a^1 \cdot \phi_j + \sum_{jj'} \mathbf{w}_a^2 \cdot \phi_{jj'} + \mathbf{w}_a^3 \cdot \mathbf{h}, \\ &= \mathbf{w}_a \cdot [\sum_j \phi_j, \sum_{jj'} \phi_{jj'}, \mathbf{h}] = \mathbf{w}_a \cdot \boldsymbol{\psi}. \end{aligned} \quad (1)$$

where the bottom equation of (1) formalizes every template as a set of parameters $\mathbf{w}_a = [\mathbf{w}_a^1, \mathbf{w}_a^2, \mathbf{w}_a^3]$ appropriately weighting the unary, pairwise and n -ary relations of G , $\boldsymbol{\psi}$. Recall that our spatiotemporal templates are extracted from unit-time segments of foreground trajectories in training. Thus, the log-likelihood in (1) is defined only for sets G consisting of unit-time trajectory segments.

From (1), the parameters \mathbf{w}_a can be learned by maximizing the log-likelihood of $\{\boldsymbol{\psi}_a^k\}$ extracted from the corresponding clusters a of training trajectories.

The log-posterior of assigning template L_a to longer temporal segments of trajectories, falling in $\tau = (t', t)$, $t' < t$, is specified as

$$\log p(L_a(\tau)|G(\tau)) \propto \sum_{k=t'}^t \log p(G^k|L_a) + \log p(L_a(\tau)) \quad (2)$$

where $p(L_a(\tau))$ is a log-normal prior that L_a can be assigned to a time interval of length $|\tau|$. The hyper-parameters of $p(L_a(\tau))$ are estimated using the MLE on training data.

4. Probabilistic Model

A parse graph is an instance of ST-AOG, explaining the event, sequence of sub-events, and human role and object label assignment. The solution of our video parsing is a set of parse graphs, $W = \{pg_i\}$, where every pg_i explains a subset of foreground trajectories, $G_i \subset G$, as

$$pg_i = \{e_i, \tau_i = [t_{i,0}, t_{i,T}], \{L(\tau_{i,u})\}, \{\mathbf{r}_{i,j}\}\}, \quad (3)$$

where $e_i \in \Delta_E$ is the recognized event conducted by G_i ; $\tau_i = [t_{i,0}, t_{i,T}]$ is the temporal extent of e_i in the video starting from frame $t_{i,0}$ and ending at frame $t_{i,T}$; $\{L(\tau_{i,u})\}$ are the templates (i.e., latent sub-events) assigned to non-overlapping, consecutive time intervals $\tau_{i,u} \subset \tau_i$, such that $|\tau_i| = \sum_u |\tau_{i,u}|$; and $\mathbf{r}_{i,j}$ is the human role or object class assignment to j th trajectory $\Gamma_{i,j}$ of G_i .

Our objective is to infer W that maximizes the log-posterior $\log p(W|G) \propto -\mathcal{E}(W|G)$, given all foreground trajectories G extracted from the video. The corresponding

energy $\mathcal{E}(W|G)$ is specified for a given partitioning of G into N disjoint subsets G_i as

$$\begin{aligned} \mathcal{E}(W|G) &\propto \sum_{i=1}^N \left[\underbrace{-\log p(\wedge_{e_i} | \vee_{\text{root}})}_{\text{select event } e_i} + \sum_u \left[\underbrace{-\log p(\wedge_{L_a} | \vee_{e_i})}_{\text{select template } L_a} \right. \right. \\ &\quad \left. \left. - \underbrace{\log p(L_a(\tau_{i,u}) | G_i(\tau_{i,u}))}_{\text{assign template}} \right] \right] \end{aligned} \quad (4)$$

where $G_i(\tau_{i,u})$ denotes temporal segments of foreground trajectories falling in time intervals $\tau_{i,u}$, $|\tau_i| = \sum_u |\tau_{i,u}|$, and $\log p(L(\tau_{i,u}) | G_i(\tau_{i,u}))$ is given by (2). Also, $\log p(\wedge_{e_i} | \vee_{\text{root}})$ and $\log p(\wedge_{L_a} | \vee_{e_i})$ are the log-probabilities of the corresponding switching OR nodes in ST-AOG for selecting particular events $e_i \in \Delta_E$ and spatiotemporal templates $L_a \in \Delta_L$. These two switching probabilities are simply estimated as the frequency of corresponding selections observed in training data.

5. Inference

Given an aerial video, we first build a video panorama and extract foreground trajectories G . Then, the goal of inference is to: (1) partition G into disjoint groups of trajectories $\{G_i\}$ and assign label event $e_i \in \Delta_E$ to every G_i ; (2) assign human roles and object labels $\mathbf{r}_{i,j}$ to trajectories $\Gamma_{i,j}$ within each group G_i ; and 3) assign latent spatiotemporal templates $L(\tau_{i,u}) \in \Delta_L$ to temporal segments $\tau_{i,u}$ of foreground trajectories within every G_i . For steps (1) and (2) we use two distinct MCMC processes. Given groups G_i , event labels e_i and role assignment $\mathbf{r}_{i,j}$ proposed in (1) and (2), step (3) uses dynamic programming for efficient estimation of sub-events $L(\tau)$ and their temporal extents τ . Steps (1)–(3) are iterated until convergence, i.e., when $\mathcal{E}(W|G)$, given by (4), stops decreasing after a sufficiently large number of iterations.

5.1. Grouping

Given G , we first use [10] to perform initial clustering of foreground trajectories into atomic groups. Then, we apply the first MCMC to iteratively propose either to merge two smaller groups into a merger, with probability $p(1) = 0.7$, or to split a merger into two smaller groups, with probability $p(2) = 0.3$. Given the proposal, each resulting group G_i is labeled with an event $e_i \in \Delta_E$ (we enumerate all possible labels). In each proposal, the MCMC jumps from current solution W to a new solution W' generated by one of the dynamics. The acceptance rate is $\alpha = \min \left\{ 1, \frac{Q(W \rightarrow W') p(W'|G)}{Q(W' \rightarrow W) p(W|G)} \right\}$, where the proposal distribution $Q(W \rightarrow W')$ is one of $p(1)$ or $p(2)$ depending on the proposal, and $p(W|G)$ is given by (4).

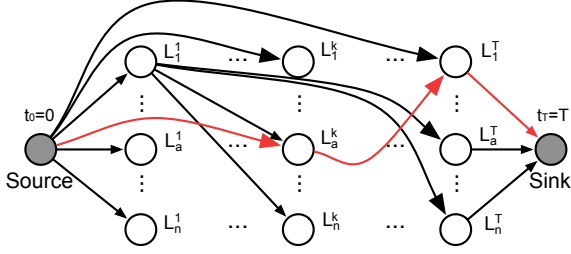


Figure 5: Our DP process can be illustrated by this DAG (directed acyclic graph). An edge between $L_{a'}^{k'}$ and L_a^k means the transition $L_{a'} \rightarrow L_a$ follows the rule defined in ST-AOG and the time interval $[t_{a'}, t_a]$ is assigned with template L_a . In this sense, with the transition rules and the prior defined in (2) (we do not consider the assignment with low prior probability), we can define the edges of such DAG. So the goal of DP is equivalent to finding a shortest path between source and sink. The red edges highlight a possible path. Suppose we find a path $source \rightarrow L_3^8 \rightarrow L_1^{20} \rightarrow sink$. This means that we decompose $[0, T]$ into 2 time intervals: $[0, 8\delta t]$, $[8\delta t, T]$, and they are assigned with template L_3 and L_1 respectively.

5.2. Human Role Assignment

Given a partitioning of G into groups $\{G_i\}$ and their event labels $\{e_i\}$, we use the second MCMC process within every G_i to assign human roles and object labels to trajectories. Each trajectory $\Gamma_{i,j}$ in G_i is randomly assigned with an initial human-role/object label $r_{i,j}$ for solution pg_i . In each iteration, we randomly select $\Gamma_{i,j}$ and change its role label to generate a new proposal pg'_i . The acceptance rate is $\alpha = \min \left\{ 1, \frac{Q(pg_i \rightarrow pg'_i)p(pg'_i|G_i)}{Q(pg'_i \rightarrow pg_i)p(pg_i|G_i)} \right\}$, where $\frac{Q(pg_i \rightarrow pg'_i)}{Q(pg'_i \rightarrow pg_i)}$ = 1 and $p(pg'_i|G_i)$ is maximized by dynamic programming specified in the next section 5.3.

5.3. Detection of Latent Sub-events with DP

From steps (1) and (2), we have obtained the trajectory groups $\{G_i\}$, and their event $\{e_i\}$ and role labels $\{r_{i,j}\}$. Every G_i can be viewed as occupying time interval of $\tau_i = [t_{i,0}, t_{i,T}]$. The results of steps (1) and (2) are jointly used with detections of large objects $\{S_i\}$ to estimate all unary, pairwise, and n -ary relations ψ_i of every G_i . Then, we apply dynamic programming for every G_i in order to find latent templates $L(\tau_{i,u}) \in \Delta_L$ and their optimal durations $\tau_{i,u} \subset [t_{i,0}, t_{i,T}]$. In the sequel, we drop notion i for the group, for simplicity.

The optimal assignment of sub-events can be formulated using a graph, shown in Fig. 5. To this end, we partition $[t_0, t_T]$ into equal-length time intervals $\{[t_{k-1}, t_k]\}$, where $t_k - t_{k-1} = \delta t$, $\delta t = 2\text{sec}$. Nodes L_a^k in the graph represent the assignment of templates $L_a \in \Delta_L$ to the intervals $[t_{k-1}, t_k]$. The graph also has the source and sink nodes.

Directed edges in the graph are established only between nodes $L_a^{k'}$ and L_a^k , $1 \leq k' < k$, to denote a possible assignment of the very same template L_a to the temporal sequence $[t_{k'}, t_k]$. The directed edges are assigned weights (a.k.a. belief messages), $m(L_a^{k'}, L_a^k)$, defined as

$$m(L_a^{k'}, L_a^k) = \log p(L_a(t_{k'}, t_k) | G_i(t_{k'}, t_k)), \quad (5)$$

where $\log p(L_a(t_{k'}, t_k) | G_i(t_{k'}, t_k))$ is given by (2). Consequently, the belief of node L_a^k is defined as

$$b(L_a^k) = \max_{k', a'} b(L_a^{k'}) + m(L_a^{k'}, L_a^k). \quad [\text{Forward pass}] \quad (6)$$

Here $b(L_a^0) = 0$. We compute the optimal assignment of latent sub-events using the above graph in two passes. In the *forward pass*, we compute the beliefs of all nodes in the graph using (6). Then, in the *backward pass*, we backtrace the optimal path between the sink and source nodes, in the following steps:

- 0: Let $t_k \leftarrow t_T$;
- 1: Find the optimal sub-event assignment at time t_k as $L_{a^*}^k = \arg \max_a b(L_a^k)$; let $a \leftarrow a^*$;
- 2: Find the best time moment in the past t_{k^*} , $k^* < k$, and its best sub-event assignment as $L_{a^*}^{k^*} = \max_{a', k'} b(L_a^{k'}) + m(L_a^{k'}, L_a^k)$; Let $a \leftarrow a^*$ and $k \leftarrow k^*$.
- 3: If $t_k > t_0$, go to Step 2.

6. Experiment

Existing Datasets. Existing datasets on aerial videos, group events or human roles are inappropriate for our evaluation. These aerial videos or images indeed show some group events, but the events are not annotated ([3, 2, 23, 22]). Most aerial datasets are compiled for tracking evaluation only [13, 12, 29]. Existing group-activity videos [8, 32, 4, 18] or social role videos [41, 9, 16, 30, 14] are captured on or near the ground surface, and have sufficiently high resolution for robust people detection. Thus, we have prepared and released a new aerial video dataset¹ with the new challenges listed in Sec. 1.2.

Aerial Events Dataset. A hex-rotor with a GoPro camera was used to shoot aerial videos at altitude of 25 meters from the ground. The videos show two different scenes, viewed top-down from the flying hex-rotor. The dataset contains 27 videos, 86 minutes, 60 fps, resolution of 1920×1080 , with about 15 actors in each video. All video frames are registered onto a reference plane of the video panorama. Annotations are provided ([37]) as: bounding boxes around groupings of people, events, human roles, and small and large objects. The objects include: 1. *Building*, 2. *Vending Machine*, 3. *Table & Seat*, 4. *BBQ Oven*, 5. *Trash*

¹Dataset can be download from <http://www.stat.ucla.edu/~tianmin.shu/AerialVideo/AerialVideo.html>

	Method	Input setting	Group	Event	Role
Baseline Var	[10] for grouping, [7] for event and role classification.	Ground-truth tracks + object annotation	77.71%	17.22%	13.98%
Baseline	Baseline method as above.	Tracking result	39.64%	16.94%	5.53%
Ours Var1	Our full model	Ground-truth tracks + object annotation	95.48%	96.38%	89.94%
Ours Var2	Our full model	Tracking result + object annotation	87.55%	54.75%	28.86%
Ours Var3	Our full model	Tracking result + group labeling	N/A	39.92%	18.71%
Ours Var4	Our model without temporal event grammar	Tracking result	40.41%	18.51%	8.69%
Ours	Our full model	Tracking result	49.47%	32.84%	18.92%

Table 1: Comparison of our method with baseline methods and variants of our approach. Our method yields best accuracy based on ground-truth bounding boxes and object labels compared to the baseline methods. Using noisy tracking and object detection results, the accuracy is limited, yet better than the baseline methods under the same condition. This demonstrates the advantages of our joint inference. When given access to the ground-truth of objects or people grouping, our results improve. Without reasoning about latent sub-events, accuracy drops significantly, which justifies our model’s ability to capture the structural variations of group events.

Bin, 6. Shelter, 7. Info Booth, 8. Box, 9. Frisbee, 10. Car, 11. Desk, 12. Blanket. The events include: 1. Play Frisbee, 2. Serve Table, 3. Sell BBQ, 4. Info Consult, 5. Exchange Box, 6. Pick Up, 7. Queue for Vending Machine, 8. Group Tour, 9. Throw Trash, 10. Sit on Table, 11. Picnic. The human roles include: 1. Player, 2. Waiter, 3. Customer, 4. Chef, 5. Buyer, 6. Consultant, 7. Visitor, 8. Deliverer, 9. Receiver, 10. Driver, 11. Queuing Person, 13. Guide, 14. Tourist, 15. Trash Thrower, 16. Picnic Person.

Evaluation Metrics. We split the 27 videos into 3 sets, such that different event categories are evenly distributed, and use a three-fold cross validation for our evaluation. Although our training and test videos show the same two scenes, we make the assumption that the layout of ground surfaces and large objects is unknown. Also, different videos in our dataset cover different parts of these large scenes, which are also assumed unknown. We evaluate accuracy of: i) grouping people, ii) event recognition, iii) role assignment. While our approach also estimates sub-events, note that they are latent and not annotated. The results are all time-averaged with the lengths of trajectories in each video. For specifying evaluation metrics we use the following notation. $G = \{G_i\}$ and $G' = \{G'_i\}$ are the sets of groups in ground-truth and inference results respectively. Γ_{ij} is the j th trajectory in i th group in ground-truth data, with duration of $|\tau_{ij}|$, group label g_{ij} , event type e_{ij} and human role r_{ij} in ground-truth. So is Γ'_{ij} in our inference. For group G_i , we call the best matched (i.e. overlapped) group in G' as M_i . For group G'_i , we call the best match group in G as M'_i . Then, precision and recall of grouping are

$$Pr_g = \sum_{G_i \in G} \left(\sum_{\Gamma_{ij} \in G_i} \mathbb{1}(M_i = g'_{ij}) \cdot |\tau_{ij}| / \sum_{\Gamma_{ij} \in G_i} |\tau_{ij}| \right) \quad (7)$$

$$Rc_g = \sum_{G'_i \in G'} \left(\sum_{\Gamma'_{ij} \in G'_i} \mathbb{1}(M'_i = g_{ij}) \cdot |\tau'_{ij}| / \sum_{\Gamma'_{ij} \in G'_i} |\tau'_{ij}| \right) \quad (8)$$

Accuracy of grouping is $F_g = 2 / (1/Pr_g + 1/Rc_g)$.

Event recognition accuracy E_e and role assignment ac-

curacy E_r are defined as

$$E_e = \sum_{G'_i \in G'} \left(\sum_{\Gamma'_{ij} \in G'_i} \mathbb{1}(e_{ij} = e'_{ij}) \cdot |\tau_{ij}| \right) / \sum_{G'_i \in G'} \sum_{\Gamma'_{ij} \in G'_i} |\tau_{ij}| \quad (9)$$

$$E_r = \sum_{G'_i \in G'} \left(\sum_{\Gamma'_{ij} \in G'_i} \mathbb{1}(r_{ij} = r'_{ij}) \cdot |\tau_{ij}| \right) / \sum_{G'_i \in G'} \sum_{\Gamma'_{ij} \in G'_i} |\tau_{ij}|. \quad (10)$$

Baselines. To evaluate effectiveness of each module of our approach, we compare with baselines and variants of our method defined in Tab. 1. For the baselines we extract the following low-level features on trajectories: shape-context like feature [8], average velocity, aligned orientation, distance from each type of large objects. All elements of feature vectors are normalized to fall in [0, 1].

Results. We register raw videos by RANSAC over Harris Corner feature points, then apply method of [12] for tracking, which is based on background subtraction [40, 33]. We also use the detector of [31] to detect buildings and cars, while other static objects are inferred in scene labeling. We do not detect portable objects, e.g., *Frisbee* and *Box*.

We evaluate our approach on both annotated bounding boxes and real tracking results. Example qualitative results are presented in Fig. 6. As can be seen, the results are reasonably good. The quantitative results are shown in Tab. 1. Confusion matrices of event recognition and role assignment are shown in Fig. 7. Additional results are presented in the supplementary material.

7. Conclusion

We collected a new aerial video dataset with detailed annotations, which presents new challenges to computer vision and complements existing benchmarks. We specified a framework for joint inference of events, human roles and people groupings using noisy input. Our experiments showed that addressing each of these inference tasks in isolation is very difficult in aerial videos, and thus provided

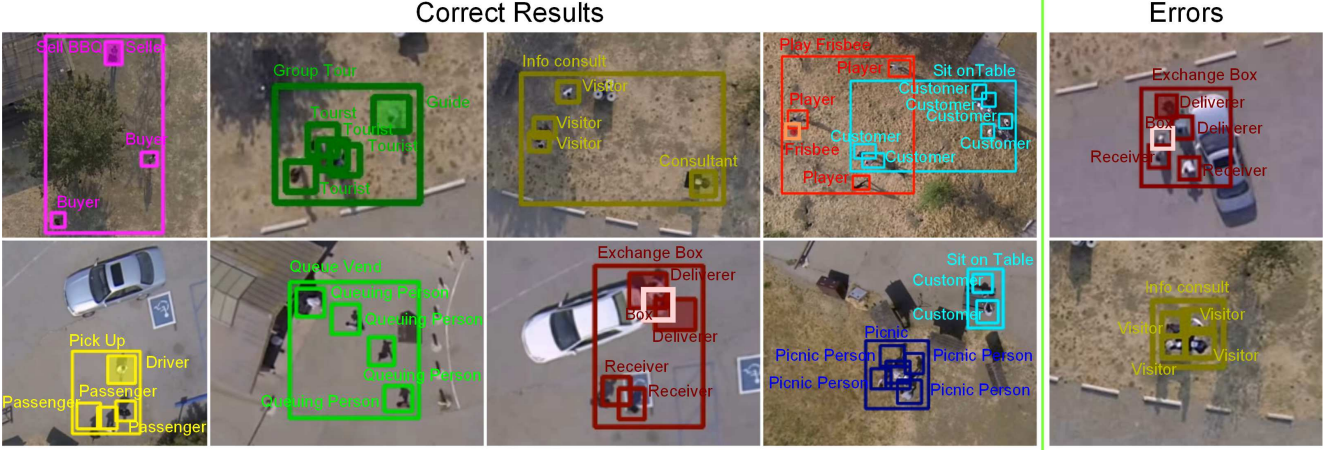


Figure 6: Visualization of results including groups (large bounding boxes), events (text) and human roles (small bounding boxes with text). In events with more than one role, we use the shaded bounding box to represent the second role; small portable objects are labeled with lighter color. From event and human role recognition, we can group people even when they are far from each other (e.g., *Play Frisbee* and *Sell BBQ*). In the top-rightmost failure example, true event *Pick Up* is wrongly recognized as *Exchange Box* because one person’s trajectory is inferred as *Box*. In bottom-rightmost failure example, our event recognition is correct, but true *Consultant* role is wrongly inferred as *Visitor* role.

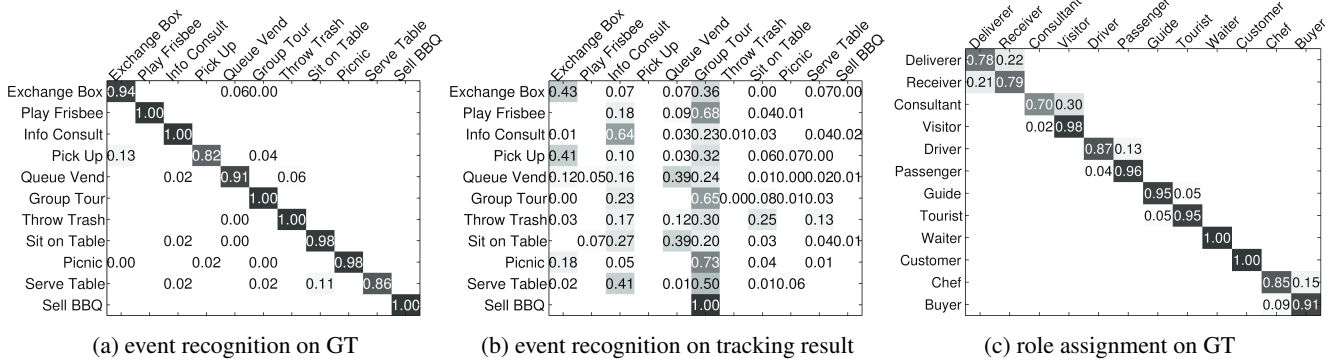


Figure 7: Confusion matrices of event recognition and role assignment result. (a) is event recognition result based on ground-truth (GT) bounding boxes and object labels; (b) is result based on real tracking and detections. From (a) and (b) we can see that *Info Consult*, *Sit on Table*, *Serve Table* cannot be easily distinguished from each other solely based on noisy tracklets. Some events (e.g. *Group Tour*) tend to be wrongly favored by our approach, especially when we do not observe some distinguishing objects. (c) is role assignment result confusion matrix within event class based on ground-truth bounding boxes and object labels. Each 2×2 block is a confusion matrix of role assignment within that event.

justification for our holistic framework. Our results demonstrated significant performance improvements over baselines when we constrained uncertainty in input features with domain knowledge.

Our model is limited and can be extended in two directions. First, we infer the function of the objects implicitly based on the group events currently. In the future, we wish to explicitly infer the functional map for a given site, in the sense that certain area corresponds to specific human activities, e.g., dinning area, parking lot, etc. Unlike appearance-based aerial image parsing [28], the spatial segmentation will be guided by the spatiotemporal characteristics of hu-

man activities. Second, similar to what [39] did for the prediction of individual intention, we would like to reason the intention of a group as another extension of our work.

Acknowledgements

This research has been sponsored in part by grants DARPA MSEE FA 8650-11-1-7149, ONR MURI N00014-10-1-0933 and NSF IIS-1423305. The authors would like to thank Dr. Michael Ryoo at JPL for the helpful discussions.

References

- [1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE TPAMI*, 34(11):2274–2282, 2012. 2
- [2] AFRL. *WPAFB*, 2009. <https://www.sdms.afrl.af.mil/index.php?collection=wpafb2009>. 6
- [3] S. Ali, V. Reilly, and M. Shah. Motion and appearance contexts for tracking and re-acquiring targets in aerial videos. In *CVPR*, 2007. 6
- [4] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S.-C. Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *ECCV*, 2012. 3, 6
- [5] B. Antic and B. Ommer. Learning latent constituents for recognition of group activities in video. In *ECCV*, 2014. 3
- [6] W. Choi, Y. W. Chao, C. Pantofaru, and S. Savarese. Discovering groups of people in images. In *ECCV*, 2014. 3
- [7] W. Choi and S. Savarese. Understanding collective activities of people from videos. *IEEE TPAMI*, 36(6):1242–1257, 2014. 2, 3, 7
- [8] W. Choi, K. Shahid, and S. Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *CVPR Workshops*, 2009. 3, 6, 7
- [9] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012. 2, 3, 6
- [10] W. Ge, T. R. Collins, and R. B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE TPAMI*, 34(5):1003–1016, 2012. 3, 5, 7
- [11] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *CVPR*, 2009. 3
- [12] Y. Iwashita, M. Ryoo, T. J. Fuchs, and C. Padgett. Recognizing humans in motion: Trajectory-based aerial video analysis. In *BMVC*, 2013. 1, 3, 6, 7
- [13] M. Keck, L. Galup, and C. Stauffer. Real-time tracking of low-resolution vehicles for wide-area persistent surveillance. In *WACV*, 2013. 1, 3, 6
- [14] S. Kwak, B. Han, and J. H. Han. Multi-agent event detection: Localization and role assignment. In *CVPR*, 2013. 3, 6
- [15] J. Kwon and K. M. Lee. Wang-landau monte carlo-based tracking methods for abrupt motions. *IEEE TPAMI*, 35(4):1011–1024, 2012. 2
- [16] T. Lan, L. Sigal, and G. Mori. Social roles in hierarchical models for human activity recognition. In *CVPR*, 2012. 2, 3, 6
- [17] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *IEEE TPAMI*, 34(8):1549–1562, 2012. 3
- [18] R. Li, P. Porfiliio, and T. Zickler. Finding group interactions in social clutter. In *CVPR*, 2013. 3, 6
- [19] L. Lin, H. Gong, L. Li, and L. Wang. Semantic event representation and recognition using syntactic attribute graph grammar. *PRL*, 30(2):180–186, 2009. 3
- [20] C. C. Loy, T. Xiang, and S. Gong. Incremental activity modelling in multiple disjoint cameras. *IEEE TPAMI*, 34(9):1799–1813, 2012. 3
- [21] R. Nevatia, T. Zhao, and S. Hongeng. Hierarchical language-based representation of events in video streams. In *IEEE Workshop on Event Mining*, 2003. 3
- [22] S. Oh et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, 2011. 6
- [23] O. Oreifej, R. Mehran, and M. Shah. Human identity recognition in aerial images. In *CVPR*, 2010. 1, 3, 6
- [24] M. Pei, Z. Si, B. Yao, and S.-C. Zhu. Video event parsing and learning with goal and intent prediction. *CVIU*, 117(10):1369–1383, 2013. 3
- [25] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You’ll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009. 2
- [26] H. Pirsiavash and D. Ramanan. Parsing videos of actions with segmental grammars. In *CVPR*, 2014. 3
- [27] T. Pollard and M. Antone. Detecting and tracking all moving objects in wide-area aerial video. In *CVPR Workshops*, 2012. 3
- [28] J. Porway, K. Wang, and S.-C. Zhu. A hierarchical and contextual model for aerial image parsing. *IJCV*, 88(2):254–283, 2010. 8
- [29] J. Prokaj and M. Gerard. Persistent tracking for wide area aerial surveillance. In *CVPR*, 2014. 1, 3, 6
- [30] V. Ramanathan, B. Yao, and L. Fei-Fei. Social role discovery in human events. In *CVPR*, 2013. 2, 3, 6
- [31] B. Rothrock, S. Park, and S.-C. Zhu. Integrating grammar and segmentation for human pose estimation. In *CVPR*, 2013. 2, 7
- [32] M. S. Ryoo and J. K. Aggarwal. Stochastic representation and recognition of high-level group activities. *IJCV*, 93(2):183–200, 2011. 2, 3, 6
- [33] A. Sobral. BGSLibrary: An opencv c++ background subtraction library. In *IX Workshop de Visão Computacional (WVC’2013)*, 2013. 7
- [34] L. Sun, H. Ai, and S. Lao. Activity group localization by modeling the relations among participants. In *ECCV*, 2014. 3
- [35] E. Swears, A. Hoogs, Q. Ji, and K. Boyer. Complex activity recognition using granger constrained dbn (gcdbn) in sports and surveillance video. In *CVPR*, 2014. 3
- [36] K. Tu, M. Meng, M. W. Lee, T. E. Choi, and S.-C. Zhu. Joint video and text parsing for understanding events and answering queires. *IEEE MultiMedia*, 21(2):42–70, 2014. 3
- [37] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *IJCV*, 101(1):184–204, 2013. 6
- [38] J. Xiao, H. Cheng, H. Sawhney, and F. Han. Vehicle detection and tracking in wide field-of-view aerial video. In *CVPR*, 2010. 3
- [39] D. Xie, S. Todorovic, and S.-C. Zhu. Inferring “dark matter” and “dark energy” from videos. In *ICCV*, 2013. 8
- [40] J. Yao and J.-M. Odobez. Multi-layer background subtraction based on color and texture. In *CVPR Workshops*, 2007. 7
- [41] J. Zhang, W. Hu, B. Z. Yao, Y. Wang, and S.-C. Zhu. Inferring social roles in long timespan video sequence. In *ICCV Workshops*, 2011. 3, 6